A Multilingual Ontology-based Lexicon for News Filtering - The TREVI project -

Hans Weigand
Infolab, Tilburg University,
email: H.Weigand@kub.nl

Abstract

Ontologies as part of a lexicon are important for many NLP tasks. In this paper, an ESPRIT project called TREVI is presented, which concerns news filtering and enrichment and draws upon a Spanish/English multilingual Lexicon. The paper describes the way the Lexicon, including the ontology, will be acquired and the methodology for building domain ontologies. In general, the approach is lexicon-driven in the sense that ontology and lexicon (vocabulary) are developed in tandem.

1 Introduction

Ontologies have been a subject of investigation in AI for several years. Loosely speaking, an ontology is a database describing the concepts in the world or some domain, some of their properties, and how the concepts relate to each other. An ontology is often organized as a classification hierarchy. Ontologies should be distinguished from domain models that are application-specific: they are intended to be used and reused in many applications, and therefore should be kept as minimal. A typical example is an ontology of the temporal domain that could be used in all applications in which reasoning with time is an issue.

Ontologies describe concepts, not the way these concepts are expressed in words in a natural language. Therefore it is usually assumed that the ontology is language-independent. There are some problems with this assumption (cf. [Weigand, 1990], [Bateman, 1993]). In the first place, there is the philosophical point that it is not possible to step outside our linguistic make-up. Concepts are shaped in the communication between members of a linguistic community. As can be illustrated by numerous examples, there are concepts that do occur in one language and not in the other, and languages differ in the way they classify the lexicon (and hence the concepts). Related to this point, it is obvious that we cannot talk about concepts without a representation. We can *distinguish* between the word and the concept, and say, for example, that two words denote the same concept, but words (or formal predicates that we create ourselves, for that matter) are indispensable as handles.

There are also some practical problems with language-independent ontologies. In particular, if we want to have a broad scope rather than a specialist domain, it seems that the best way to start is with available machine-readable dictionaries, such as WordNet ([Miller, 1995]), although it is clear that such a dictionary has a language bias. The best approach seems to be developing the ontology and the lexicon(s) in parallel. In Cyc, for example, there is not always a clear separation between knowledge of English words (the lexicon) and knowledge of concepts ([Mahesh, Nirenburg et al, 1996]), resulting in ad-hoc solutions and a lack of genericity.

In this paper, I will give a short overview of a recently started NLP project our group is involved in, in which ontologies and lexicons have to be developed. Next, I will describe the approach that we want to take. At this moment, we cannot report yet on actual experience, but more material will be available at the time of the workshop.

2 TREVI: News filtering and enrichment

TREVI is an ESPRIT project (#23311) started in January 1997 in which Tilburg University is responsible for the Lexicon Management System and its contents. The project is managed by ITACA (Rome, Italy).

The TREVI (Text Retrieval and Enrichment for Vital Information) Project aims at offering a solution to the problem of "information overflow", i.e. the difficulty experienced both by large companies and SMEs (Small and Medium-size Enterprises) in extracting useful information from large amounts of data coming from the numerous electronic textual information services available at local or global level (Internet, proprietary networks,

subscription services, World Wide Web, and more).

The key result of the TREVI Project will be a set of software tools (the TREVI Toolkit) representing a substantial improvement in the management of distributed textual information sources. The TREVI Toolkit will not rely on simple text-based search tools; it will rather combine concept-based search and active data mining techniques to enrich online input text streams by providing indexation, abstraction, smart correlation with data and knowledge sources, compilation into electronic publication formats, and subscription capability on the results through communication services (for example, HTML document servers on WWW).

The work will be approached starting from a market understanding based on actual user requirements. The project is driven by a consortium of software development enterprises and users with a need to heavily exploit information technology to retain their competitiveness in the evolving market. User partners are all involved in business areas, ranging from medicine to the information service industry, which deserve fast and efficient management of large amounts of textual information. They will provide demonstrators based on their own business cases, which will allow to test the effectiveness of the Projects' basic concepts and show the flexibility of the TREVI Toolkit architecture in real-world, critical scenarios.

The languages supported by TREVI include Spanish and English. It has been decided therefore to build one multilingual Lexicon Management System (LMS) with a shared (language-independent) semantic part (the concept base) and separate language-dependent morphosyntactic parts (the lexicons). The LMS has to cover both domain-independent lexical semantics and (a limited number of) specific domains.

The main functions of the LMS are (1) support of the parsing process and (2) support of the subject identification. For the former, the lexicons are the most important, whereas for the latter task, the concept base is crucial. Subject identification includes the matching with user profiles that have been set up with the same concept base, and implies the computation of the semantic distance between the document representation (in terms of concepts) and the user profile (also in terms of concepts).

2.1 Word sense disambiguation

A central problem in TREVI, as in NLP in general, is word sense disambiguation. Subject matching on the basis of concepts requires that the concepts are first identified correctly. However, words are ususally ambiguous. The basis problem is then to select the combination of word senses in a sentence or text that best fits the overall meaning and context of the sentence. Word sense ambiguities can be classified in three types [Mahesh, 1996]:

1. One sense fits the context, and other ones are anomalous;
2. Two or more senses are acceptable, but one is better
3. All senses are anomalouos, but one must be chosen nevertheless

For word sense disambiguation, it is essential to have information about selection restrictions (frame role restrictions). And concepts must be organized in a taxonomy so that constraints can be stated concisely. For example, the verb "sell" takes a person or organization as agent. The taxonomy should include that a bank is an organization, so that "bank" can fit in the agent role of "sell".

What is also important is that the number of concepts is kept low. For example, the word "bank" has different senses. In WordNet, it has, among others, the sense of "financial institution" and the sense of "bank building". It is clear that these senses are closely related; the latter can be viewed as a projection of the first (cf. section 5). At that moment, it is the question whether the difference is important enough for the application at hand to maintain it: if a TREVI user is interested in "banks", he should not get messages about river banks of course, but messages about bank buildings (for example, "a new bank has been opened yesterday in Bilbao") seem to be close enough to include them.

3 Ontologies: a lexicon-driven approach

Since the LMS has the task to support the parsing process in TREVI, it should ideally provide full lexical semantics for a substantial vocabulary (say, 100,000 words). Since this is not feasible with the limited resources that we have, a more subtle approach is needed. This approach makes a distinction between three sets of words (and concepts, respectively): the core, the crowd, and the chosen:

1. "core" containing the basic concepts, that is, not just very general categories like CONCRETE but the basic classifications (natural kinds), such as ANIMAL, PERSON, COUNTRY, and also HORSE, MAKE, EAT, etc (cf. [Rosch, 1977] [Vossen, 1995]). Starting point for this set is the set of most familiar concepts according to WordNet; this set will be analyzed and augmented by hand to ensure language-independence and minimality. On the basis of some experiments with LDOCE, Vossen (1995) has estimated the number of basic-level (nominal) concepts at about 11,000. We expect the core to contain about 2,000 items, taking only the ones most frequently used in the TREVI corpora, but it can grow during use. A similar number of basic level action concepts will be included. Action concepts (corresponding with verbal lexemes) will be augmented

with role semantics ([Dik, 1989], [Jackendoff, 1983]). The core will be language-independent and have expressions in both English and Spanish.

2. "crowd" containing all other concepts that are needed for lexical semantics. For English, these concept descriptions will be based on (English) WordNet synsets. Ideally, these concepts would all be expressed in terms of the core ontology, but this will not be feasible for practical reasons.

   For Spanish, the crowd will be built up from other resources. In principle, there will be no sharing between the two languages beyond the common core. In this way, discussions about translation equivalence are minimized, and the Spanish and English concept base can be developed in parallel and independently.

   The conceptual content of the crowd will be more limited than that for the core. Due to our limited resources, many lexical entries will not even get a link to a concept. However, the LMS gives full support for extensibility, so that in practice, the number of links between lexical entries and concepts will grow steadily in a way determined by the application.

3. "chosen" containing more elaborate conceptual information for specific domains (the ones relevant for the project). Domain concepts are linked to the core, and, where necessary, to the crowd, since no domain is an island. The structure of domain ontologies is described below.

The advantages of splitting core and crowd are: (1) the "core" provides us with a comprehensible (cognitively relevant) way of structuring the concept set; while (2) language-dependent nuances are not excluded, but also not analyzed deeper than strictly necessary for the requirements of the application. When in the future, lexicons for more languages become available (and are required in the system), they can be added to the LMS with minimal effort: we require only the core to be shared. Note that the core is not fixed, and the addition of another language may prompt extensions.

Note that we do not analyze all concepts in the same depth. Since there is no principled boundary between ontological and general-world (encyclopedic) knowledge, the boundary will be arbitrary and "situated" (in the sense of [Mahesh, 1996]) anyway.

4 Domain analysis

Ontologies, like terminologies in the past, are typically thought of as taxonomic structures. In our experience with domain modelling so far, taxonomies are less central than they appear to be and often much more arbitrary than the analyst wants to acknowledge. In TREVI, we will take a different approach in which taxonomies are only secondary.

In accordance with principles of Object-Oriented Analysis (e.g. [Kristen, 1994]), *actions* (or events) are taken as central. In OO, an object type is determined by the actions, or methods, that it can perform, not on the basis of its structure. The actions, expressed by action verbs, correspond to practices in the domain that do not change very much over time. On the other hand, the terms that label the agents and objects involved can be changed easily. So whether temporary workers are called employees or not, is something that can be changed over night. But the actions that they perform, and the actions the organization performs on them, remain the same. The domain analysis methodology we want to adopt contains roughly the following steps: model actions, resulting in an Action Ontology; analyze terms, resulting in an Object Ontology; analyze derived terms, resulting in a Terminology.

4.1 Action modelling

Find the relevant actions in the domain, starting with the action verbs encountered in texts (explicit or hidden, as in nominalizations). For each action, determine the role or frame structure (agent, patient etc) and the selection restrictions. The selection restrictions should be filled by basic concepts of the core ontology. They should not be role names (so "person" instead of "employee").

The total set of essential actions and the roles makes up a conceptual network comparable to an Entity Relationship diagram. We call it the Action Ontology. The actions are categorized according to prototypical event structures, such as TRANSFER, TRANSFORM, TRANSPORT and ACT ON ([Jackendoff, 1983]). No more conceptual information is defined yet.

4.2 Term analysis

Analyze the list of terms one by one. A term is a simple or compound Noun Phrase expressing some kind of entity type. Nominalizations (disguised actions) and reified properties are not taken into account: the former have been treated by the Action Model, the latter will be treated in step (3). We also do not include instances here, such as country names, but they will be included in (a special section of) the LMS. For polysemous terms, the procedure is applied to each sense (although it should be attempted to unify different senses under one prototype wherever possible).

A distinction is made between the following term classes:

names: terms that express basic level concepts. Usually, a name will have been encountered in the Action

Model. If not, it is likely to be irrelevant. However, the Action Model only described the actions that the entity is involved in. At this stage, we define for each name a concept frame containing prototypical information, such as a TELIC role (KNIFE telic CUT) or more complex roles expressed by an action. Such prototype information seems most appropriate to "names", because they correspond to rich concepts. For the other term classes below, the relevance of prototype information remains to be considered.

roles: terms that express a role of an entity in some action, for example, EMPLOYEE and EMPLOYER. Roles can be defined in terms of the action model, for example: an EMPLOYER isa (PERSON or ORGANIZATION) who does EMPLOY a PERSON (in this case, the term is a superordinate and a role), or DIRECTOR isa PERSON who does DIRECT an ORGANIZATION (in this case, the term is is a role and a subordinate). Roles are completely defined by the relationship they express (in this case, EMPLOY and DIRECT, respectively). Since these relationships are made explicit in the Action Model, we can simplify the Term Model by sorting the roles out.

Roles are distinguished in several OO methods (KISS [Kristen, 1994], NIAM), but not in the work of [Mahesh, 1996].

superordinates/generalizations: terms that express a specific property (or capability) that is shared by a number of (basic) concepts. An example is VEHICLE as a generalization of CAR, SHIP etc, or LIQUID as a generalization of WATER, WINE, BLOOD etc). Generalizations are specified by means of the property they express, and this property is then inherited to its hyponyms.

subordinates/specializations: terms that classify basic level concepts according to some property. For example, PERSON can be specialized to MAN and WOMAN according to the property GENDER. Specializations are specified by means of the basic level concept they start from and the properties that they express. Note that roles are treated separately. Subordinates inherit properties of the basic concept they are attached to, but such prototypical information can be overruled.

component/part: terms that are defined by means of a PART-OF relation to a basic (from the point of view of aggregation) concept. For example, NOSE is defined as PART-OF a PERSON.

group: terms that are defined as an aggregation of basic concepts. For example, TEAM is defined as a GROUP-OF PERSON.

The total set of essential object concepts can be organized in the form of two hierarchies, or tree diagrams: a hyponymy hierarchy and a meronymy hierarchy. In such a hierarchy, the distinction between basic and non-basic is blurred. However, the distinction can still play a role, for example, in the presentation of query results. When a user wants to know what a SCHNAUTZER is, the reply should be that it is a kind of DOG, and not just list all hyperonyms.

4.3 Terminology

The Terminology is defined here as the set of derived or analytical terms. Analytical terms express some property determined by some definition. For example, AGE being defined as the number of years after the BIRTH-event, or NET INCOME being defined as the GROSS SALARY minus taxes, where GROSS SALARY in turn is defined as the MONEY EARNED-BY the PERSON.

5 Dynamics: a generative ontology

Just as the formal part of a Lexicon contains generative rules, such as compounding, diminutive formation etc, so does the conceptual part, including the ontology. An example is the *projection* operator that maps an EVENT concept to a concept representing a participant of the EVENT (e.g. EMPLOY to EMPLOYER). Another one is *composition*, where a concept frame is specialized by replacing one of the role fillers by a more specific concept (e.g. RAISE(ag)(pt) + POTATO ==> RAISE(ag)(pt POTATO)), and *selection* (e.g. RAISE(ag)(pt POTATO) + (loc HOLLAND) ==> RAISE(ag)(pt POTATO)(loc HOLLAND)). In principle, the recognition of such operations make the explicit stipulation of the generated concepts superfluous. The ontology should be able to generate these concepts on the spot. Depending on whether we prefer minimality or comprehensiveness, we either exclude or include generated concepts in the ontology. When we do not include them in the ontology, it is still possible to include them in the concept base and to have the corresponding words in the (English or Spanish) lexicon. In this way, the ontology is not larger than strictly necessary. The same mechanism can be used for defining the terminology (see above), thus keeping the ontology as concise as possible. This leads to a distinction within the domain model similar to the one between core and crowd above: the ontology is the core of the domain model, and the terminology the crowd. The ontology is defined relationally (the meaning is in the links between the concepts), whereas the terminology is defined analytically (the meaning is a definition).

The generative operations that we identified above (projection, composition and specialization) correspond to the basic graph operators defined by ([Sowa, 1984]). On formal grounds one could argue that these are sufficient, but we leave it to empirical verification to decide on this. Linguistic evidence on productive word formation (the generative rules of the lexicon) can provide this empirical basis.

6 Conclusion

In this paper, I have sketched a rough outline of the approach that we want to follow in TREVI for the construction of a multilingual Lexicon Management System. In the LMS, ontology and dictionary (-ies) are developed in tandem. We have described both the way we want to set up a general concept base, with a core ontology, in parallel to a multilingual lexicon, and the way we want to set up in-depth ontologies for specific domains. Ultimately, the concept base could be replaced by a (large) set of domains, but we don't expect that to happen in the limits of this project.

Acknowledgments

The TREVI project is supported by the European Commission from 1997 till 1999. For more information, see: http://www.itaca.it/itaca/it/Trevihom.htm.

References

[Bateman, 1993] J. Bateman Ontology construction and natural language In: N. Guarino, R. Poli (eds), *Proc. of the Int. Workshop on Formal Ontology in Conceptual Analysis and Knowledge Representation* Padova, March 1993.

[Dik, 1989] S.C. Dik *The Theory of Functional Grammar* Foris, Dordrecht, 1989.

[Jackendoff, 1983] R. Jackendoff *Semantics and Cognition* MIT Press, Cambridge, 1983.

[Kristen, 1994] G. Kristen *Object-Orientation: The KISS-method: From Information Architecture to Information Systems* Addison-Wesley, 1994.

[Mahesh, Nirenburg et al, 1996] K. Mahesh, S. Nirenburg et al An Assessment of Cyc for Natural Language Processing MCCS-96-302, New Mexico State University, 1996.

[Mahesh, 1996] K. Mahesh Ontology Development for Machine Translation: Ideology and Methodology MCCS-96-292, New Mexico State University, 1996.

[Miller, 1995] G.A. Miller WordNet: A Lexical Database for English *Communication of the ACM*, 38(11), Nov 1995.

[Rosch, 1977] E. Rosch Classification of real-world objects: origins and representation in cognition In: P.N. Johnson-Laird and P.C. Wason (eds), *Thinking: readings in cognitive science* Cambridge Univ Press, 1977.

[Sowa, 1984] J. Sowa *Conceptual Structures: Information Processing in Mind and Machine* Addison-Wesley, 1984.

[Vossen, 1995] P. Vossen Grammatical and Conceptual Individuation in the Lexicon Ph.D. Thesis, Univ of Amsterdam, 1995.

[Weigand, 1990] H. Weigand *Linguistically Motivated Principles of Knowledge base Systems* Foris, Dordrecht, 1990.